doi:10.6041/j.issn.1000-1298.2020.05.023

基于 TextRank 和簇过滤的林业文本关键信息抽取研究

陈志泊[1]　李钰曼[1]　许　福[1]　冯国明[2]　师栋瑜[3]　崔晓晖[1]

(1.北京林业大学信息学院,北京 100083;2.中国联合网络通信集团有限公司,北京 100033;
3.中国电信系统集成有限责任公司,北京 100035)

摘要:目前,获取林业文本关键信息存在 2 个问题:关键信息获取主要从关键词角度考虑,忽略了词语的信息类型;网络上的林业文本没有统一的记述结构,词语信息类型提取困难。为此,本文提出了基于改进 TextRank 和簇过滤的林业文本关键信息抽取方法,以“关键词 + 信息类型”两部分表示文本关键信息。首先,抽取关键词并进行 Word2Vec 向量化,然后通过构建融合词语特征值、边权值的图模型对 TextRank 进行改进,对经迭代收敛得到的稳定图进行归并聚类形成簇;然后,设计簇品质评价公式进行簇过滤,再次应用 TextRank 形成最终簇集合;最后,对簇进行信息类型标注。对于测试文本,通过比较关键词向量和簇心向量的距离获得词语的信息类型,将信息类型与关键词结合得到文本的关键信息。基于 2 000 篇与林业政策新闻相关的林业文本进行实验,最终簇集合的紧密度为 0.968 0,间隔度为 0.057 2,综合评价指标为 0.887 1;对其中 400 篇文本进行关键词人工标注,将本文关键词抽取方法与 TextRank、TF - IDF 等 6 种算法进行比较,结果表明,本文方法在 MRR、Bpref、准确率和综合评价指标上均获得了较好的效果,说明本文方法在提取林业文本关键词方面具有优势。

关键词:林业文本;关键词抽取;TextRank;簇过滤;信息类型

中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1000-1298(2020)05-0207-08　　**OSID**:

Key Information Extraction of Forestry Text Based on TextRank and Clusters Filtering

CHEN Zhibo[1]　LI Yuman[1]　XU Fu[1]　FENG Guoming[2]　SHI Dongyu[3]　CUI Xiaohui[1]

(1. *School of Information Science and Technology*, *Beijing Forestry University*, *Beijing* 100083, *China*
2. *China United Network Communications Group Co.*, *Ltd.*, *Beijing* 100033, *China*
3. *China Telecom System Integration Co.*, *Ltd.*, *Beijing* 100035, *China*)

Abstract: There are two main problems in obtaining key information of forestry text, firstly, the key information is mainly considered from the perspective of keywords, and the information types of words are neglected; secondly, there is no unified description structure for forestry text on the Internet, which makes it difficult to extract word information types. Through combining the two characteristics of “keywords + information types”, a method about forestry text key information extraction was proposed based on inproved TextRank and clusters filtering. The main contents were as follows: the first step was to extract the text keywords according to the keywords extraction formula. The second step was to characterize the keywords with Word2Vec vectorization. The third step was to improve the TextRank algorithm, mainly by merging the word features and introducing the edge weights to construct the graph model of the text. The fourth step was to obtain the stable graph structures through iterative convergence, and then merged them to form clusters. And the clusters's quality was evaluated from three aspects: the uniformity of elements distribution, the size of the clusters, and the universality of the clusters. The fifth step was to form the final clusters' set in combination with the TextRank algorithm. The final step was to label the final clusters about information types. The data used in the experiments were 2 000 forestry texts related to forestry policies and news. The experimental results showed that compactness of the final clusters' set was 0.968 0, the separation of the final clusters' set was 0.057 2, and the F1-measure of the final

收稿日期:2019 - 12 - 31　修回日期:2020 - 02 - 23

基金项目:国家自然科学基金项目(61772078)和北京林业大学热点追踪项目(2018BLRD18)

作者简介:陈志泊(1967—),男,教授,博士生导师,主要从事大数据技术和计算机软件与理论研究,E-mail:zhibo@bjfu.edu.cn

通信作者:许福(1979—),男,教授,主要从事软件工程和林业遥感技术研究,E-mail:xufu@bjfu.edu.cn

clusters' set was 0.887 1. It showed that the information types of the clusters can be clearly marked. For a text's keywords, their information type was obtained by calculating the cosine similarity of the keywords' vector and the clusters' heart. The combination of keywords and information types constituted key information of a foresty text. Meanwhile, manually labeled 400 texts, comparing with the six algorithms such as TextRank, TF - IDF, this method achieved the better results in MRR, Bpref, accuracy, and F1-measure. It showed that this method had advantages in extracting forestry text keywords.

Key words: forestry text; keywords extraction; TextRank; clusters filtering; information types

0 引言

随着互联网与人工智能技术的飞速发展,我国传统林业也逐步向“智慧林业”迈进。对于网络上数量呈爆发式增长的林业文本来说,如何节省阅读时间、从中准确获取与林业领域有关的信息具有重要的研究意义[1-2]。

文本关键信息应包含关键词和信息类型。目前大多数的林业文本并没有标注关键词,早期的关键词抽取是通过人工标注、借助人类的专业知识完成的,工作任务十分繁重。随着计算机技术的发展,借助计算机程序抽取关键词成为更好的选择[3-6]。关键词抽取主要分为有监督和无监督两类[7]。由于有监督算法标注成本高,且存在过拟合的问题,近年来,无监督关键词抽取算法得到广大科研人员的青睐。常见的无监督关键词抽取方法有 3 种:基于统计特征[8-9]、基于词图模型[10-12]和基于主题模型[13]的关键词抽取。基于统计特征的关键词抽取算法[14-15]将文本词语的统计信息记为特征信息,如词频特征、逆文档频率特征、长度特征、位置特征等,再对特征信息进行相应的量化处理,最后抽取出文本关键词。其缺点是忽略了词语之间的相互关系,效果有时并不理想。基于主题模型的关键词抽取算法[16-17]认为每个文本都对应着一个或多个主题,而每个主题都会有相对应的词分布,通过分布信息得到文本与词的关联情况,进而得到文本关键词,以 LDA 隐含主题模型为经典代表。其缺点是模型需要大量的数据训练,对于内容较短的文本不敏感,且计算复杂度较高,所以提取效果有时并不理想。基于词图模型的关键词抽取算法通过融合词语特征信息达到优化提取效果的目的,是目前应用最广的无监督提取方法[18]。因此,融合词语的特征信息[19-23]、优化词图模型、提高抽取效果具有重要的研究价值。仅抽取关键词不能完整且直观地表达文本内容,因而需要借助信息类型来完善。对于词语的信息类型,如果文本有严格的记述特征,则可以通过分析记述结构进而抽取到相应的属性[24-27],但大多数文本没有良好的记述结构,故采用此方法相对困难。因此,如何确定词语的信息类型具有现实意义。

本文采用合理的公式抽取关键词,通过改进 TextRank 算法、归并聚类、簇过滤等,获取到高品质的词语集合,进而进行信息类型的判定,将关键词和信息类型结合,实现对林业文本的关键信息抽取。

1 实验方法

1.1 相关技术原理

1.1.1 词语特征

关键词抽取是指从文档中获取有代表性的词语,用以反映文档的主题和核心内容。衡量词语的重要性,不能从单一角度考虑。通常用词频特征、位置特征等[28-29]特征来衡量。文献[30]已对基于词频-逆文档频率特征、长度特征、词语首次出现的位置特征以及词跨度特征等 4 个方面关键词抽取公式进行了相关研究,基于此,本文在该基础上加入标题特征,如果词语出现在标题中,标题特征值记为 1.5,反之记为 1。词语综合权重值计算公式为

$$V = W_{tf}W_{idf}\mathrm{lb}l(1+\mathrm{e}^{-s})\left(1+\frac{t-s}{n}\right)\theta \tag{1}$$

式中 W_{tf}——词频 l——词长

W_{idf}——逆文档频率

s——词语首次出现的位置

t——词语最后一次出现的位置

n——文本词汇总数

θ——标题特征值

其中 $W_{tf}W_{idf}$为词频-逆文档频率特征,$\mathrm{lb}l$ 为长度特征,$1+\mathrm{e}^{-s}$为词语首次出现的位置特征,$1+(t-s)/n$ 为词跨度特征。

1.1.2 相似性度量

将关键词用 Word2Vec 向量化表征,根据向量之间的相似度聚为若干簇。相似性度量方法有:欧氏距离、曼哈顿距离、切比雪夫距离、闵可夫斯基距离、马氏距离、余弦相似度、汉明距离等。本文通过设置阈值,计算向量间的余弦相似度对向量进行归并聚类,向量余弦相似度 $W_{\cos}$计算公式为

$$W_{\cos} = \frac{\boldsymbol{X}\boldsymbol{Y}}{\|\boldsymbol{X}\|\|\boldsymbol{Y}\|} \tag{2}$$

式中　$\boldsymbol{X}$、$\boldsymbol{Y}$——任意两个不同的向量

$\|\boldsymbol{X}\|$、$\|\boldsymbol{Y}\|$——$\boldsymbol{X}$、$\boldsymbol{Y}$ 向量的模

1.1.3　TextRank 算法

TextRank 算法是一种用于文本的基于图的排序算法，通过对图结构的迭代计算实现词语的重要性排序[31-33]。优点是不需要事先对文档进行相关的学习训练。基本原理如下：

设 $G(V,E)$ 是由给定文本的词汇所构成的图结构，V 为图节点集合，E 是图边集合。对于文本中的任一 V_i，基于 TextRank 算法得到权值 W_i 计算式为

$$W_i = 1 - d + d \sum_{V_j \in \mathrm{In}(V_i)} \frac{w_{ji}}{\sum_{v_k \in \mathrm{Out}(V_j)} w_{jk}} W_j \tag{3}$$

式中　d——阻尼系数，取值为 0 ~ 1

$\mathrm{In}(V_i)$——指向节点 V_i 的所有节点的集合

$\mathrm{Out}(V_j)$——节点 V_j 指向的所有节点的集合

w_{ji}、w_{jk}——节点 V_j 到节点 V_i、V_k 的边的权重

W_j——节点 V_j 的权值

但最初 TextRank 算法在应用时，忽略了词语本身的特征信息，使各节点初始值均等，且节点权重均匀转移。

因此，本文在原有 TextRank 算法的基础上进行改进，考虑词语特征，将由关键词抽取公式计算得到的综合权值作为节点的初始值，并用词语向量间的余弦相似度作为边的初始值，构建带权无向图结构，此时对于文本中的任一 V_i，权值 W_i 计算公式为

$$W_i = 1 - d + d \sum_{V_j \in \mathrm{join}(V_i)} \frac{w_{ji}}{\sum_{v_k \in \mathrm{join}(V_j)} w_{jk}} W_j \tag{4}$$

式中　$\mathrm{join}(V_i)$——节点 V_i 相连的所有节点的集合

1.1.4　簇过滤

对于向量归并聚类形成的簇来说，需要通过合理的品质评价指标对簇的品质进行过滤。本文从簇元素分布的均匀性、簇的规模、簇的普适性 3 个角度考虑，设计簇品质评价公式，通过对相关参数的调整，过滤得到品质比较好的簇集合。

(1)簇元素分布的均匀性(Balance)

簇中元素分布均匀性指标 B 的计算公式为

$$B = \frac{A_v}{1 + S_t} \tag{5}$$

式中　A_v——簇中节点元素权值的平均值，反映数据的集中性特征

S_t——簇中节点元素权值的标准差，反映数据的波动性特征

标准差 S_t 越小，说明元素的权重分布越均匀，说明簇中有用的元素越多。标准差 S_t 相同时，需要借助平均值 A_v 来判定。当 S_t/A_v 越小，元素分布越均匀，依据取倒数且分母不为零的原则设计公式，得出 B 值越大，簇元素分布越均匀。

(2)簇的规模(Scale)

簇的规模 S 指簇中所含元素的数量，计算公式为

$$S = \frac{1}{1 + \mathrm{e}^{-\mu}} - 0.5 \tag{6}$$

式中　μ——簇中元素的个数

当簇中元素越多，即 μ 越大，S 越大，说明簇的规模越大。

(3)簇的普适性(Universality)

普适性指标 U 是指簇中元素来源的文章数，计算公式为

$$U = \frac{1}{1 + \mathrm{e}^{-N_0 + 1}} - 0.5 \tag{7}$$

式中　N_0——簇中元素来源的文章数

簇中元素来源的文章数越大，即 N_0 越大，U 越大，说明簇的普适性越好。

(4)簇品质(Quality)评价公式

从簇元素分布的均匀性、簇的规模、簇的普适性 3 个角度考虑簇的品质 Q，计算公式为

$$Q = \lambda_1 BS + \lambda_2 U \tag{8}$$

式中　λ_1、λ_2——参数

Q 由两部分组成：簇的总体水平和簇的普适性，通过调节参数，权衡各指标所占的权重。说明在不同规模下，均匀程度的增大对簇的品质的提升是不同的；规模越大，提升越大。所以簇品质对均匀程度的偏导为规模的单增函数，可得

$$\frac{\mathrm{d}Q}{\mathrm{d}B} = f(S) \tag{9}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}S} = g(B) \tag{10}$$

式中函数 f、g 均为单增函数。为方便计算，设

$$f(x) = g(x) = \lambda_1 x \tag{11}$$

因此，簇的总体水平由簇元素分布的均匀性和簇的规模两部分组成，可记为 $\lambda_1 BS$。

1.2　提取流程

针对数量庞大的林业文本，采用“关键词 + 信息类型”的表示方式，提出基于改进 TextRank 和簇过滤的林业文本关键信息抽取方法，通过合理的方式对林业文本进行关键信息抽取。提取流程如图 1 所示，具体步骤为：

(1)林业文本预处理，包括引入领域词典对文本进行分词、引入停用词表对文本进行去停用词等操作。

(2)依据关键词抽取公式，抽取文本综合权值

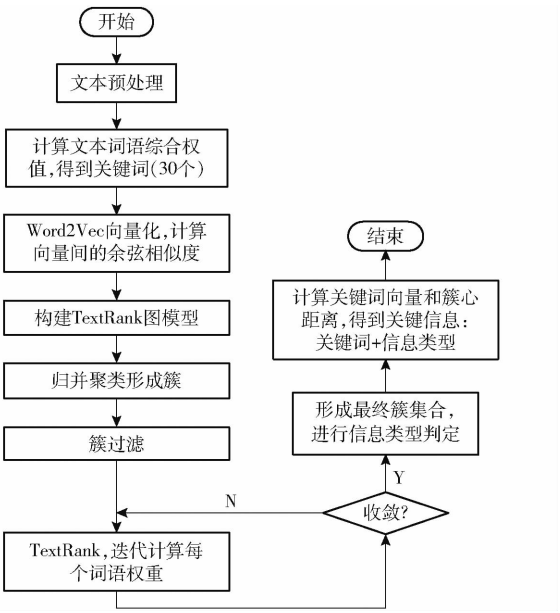


图 1 关键信息提取流程

Fig. 1 Key information extraction process

排名前 30 的词语,部分结果如图 2 所示。

[林权制度改革, 漯河市, 强力, 四项措施, 林权制度, 圆满完成, 目标管理, 漯河, 市政府, 促进会, 明确责任, 改革领导小组, 上下联动, 齐抓共管, 成员单位, 各县区, 具体操作, 河南省, 阶段性, 不合格, 基本合格, 动员, 奖励, 奖惩, 工作机制, 攻坚战, 路沟, 上档次, 林改, 考核指标]
[农村改革, 全面建设小康社会, 分田, 农村, 第一村, 重大突破, 联产承包, 农村土地, 田村, 承包经营权, 两者, 又一次, 相似之处, 生产力, 比较简单, 考虑, 中国农村, 安徽凤阳, 制度, 某种意义, 退避三舍, 双重性, 手印, 福建永安, 新农村建设, 组成部分, 发展潜力, 责任制, 分山, 潜力]
[交易平台, 花木, 肥西, 苗木花卉, 中部, 电子交易, 平台, 蒇汇, 首个, 转型发展, 数十万亩, 产业升级, 三足鼎立, 正式启动, 城及, 十八届三中全会, 农惠农, 资产交易, 交易额, 种植基地, 乃至, 亿, 合肥, 我省, 亿, 月, 西有, 温江, 全国领先, 中国]
[集体林权制度, 改革试点, 江苏省, 立足于, 江苏省政府, 小组办公室, 培训班, 实施意见, 铜山县, 溧阳市, 如皋市, 盱眙县, 阜宁县, 宿迁市, 宿城区, 南京市, 浦口区, 试点工作, 高邮市, 姜堰市, 灌南县, 纪委书记, 党委委员, 试点县, 邱昱东, 下称, 精神实质, 深刻领会, 邱昱, 东到]

图 2 部分抽取结果

Fig. 2 Part of extraction results

(3)对抽取的关键词用 Word2Vec 向量化表征,并计算两两向量的余弦相似度。

(4) 设置阈值,向量余弦相似度大于阈值的两个词之间连线,间距小于阈值的 2 个词之间不连线,以相似度为边的权值,以步骤(2)计算出的权值作为节点的初始权值,进而构造图模型,应用 TextRank 算法,得到了综合考虑词与词关系的关键词最终权值。

(5)利用图结构中词语的节点值和相应的词向量进行加权求和,得到图中心。计算两两图中心的余弦相似度,设置阈值,余弦相似度大于阈值的图进行合并,归并聚类得到初始簇。

(6)对簇中的节点值进行标准化处理,依据设计好的簇品质评价公式,对初始簇进行品质评价,设置阈值,进行过滤操作。

(7)对过滤后的簇应用 TextRank 算法,经过迭代收敛得到最终簇集合,对最终形成的簇集合进行信息类型的判定。

(8)计算关键词向量和各簇心之间的余弦相似度,通过比较,得到关键词的信息类型。最终得到文本的关键信息:关键词 + 信息类型。

2 实验及方法验证

2.1 实验环境

本文所提出的算法模型采用 Python 编程实现,本实验所有的模型训练计算机环境主要参数为 Intel Corei5 -8250U CPU @ 1.6 GHz 1.80 GHz,内存为 8.00 GB。

2.2 实验数据

本文所采用的实验数据为与林业政策和新闻相关的文本,数据分别来自中国林业新闻网、林业信息网、林业产业网等林业相关网站,经数据预处理后共 2 000 篇,其中 400 篇文本进行了关键词人工标注。

2.3 评价指标

2.3.1 聚类评价指标

所采用的评价指标有 3 个:紧密度、间隔度、聚类综合评价指标。

(1)紧密度(Compactness,CP)

每一个簇中各元素到簇心的平均距离越小,说明聚类效果越好。实验选用向量的余弦距离,因此 CP 越大,说明聚类效果越好。

$$C_p = \frac{1}{k}\sum_{i=1}^{k}\overline{C}_{pi} \tag{12}$$

其中

$$\overline{C}_{pi} = \frac{1}{\|\boldsymbol{\Omega}_i\|}\sum_{\boldsymbol{x}_i\in\boldsymbol{\Omega}_i}\|\boldsymbol{x}_i - \boldsymbol{w}_i\| \tag{13}$$

式中 C_p——紧密度

$\boldsymbol{x}_i$——簇中第 i 个关键词向量

$\boldsymbol{w}_i$——第 i 簇的簇心向量

$\boldsymbol{\Omega}_i$——第 i 簇的关键词集合

k——簇的个数

(2)间隔度(Separation,SP)

各簇中心两两之间的平均距离越远说明簇间聚类效果越好。实验选用向量的余弦距离,因此间隔度越小,说明聚类效果越好。

$$S_p = \frac{2}{k^2 - k}\sum_{i=1}^{k}\sum_{j=i+1}^{k}\|\boldsymbol{w}_i - \boldsymbol{w}_j\|^2 \tag{14}$$

式中 S_p——间隔度

$\boldsymbol{w}_j$——第 j 族的簇心向量

(3)聚类综合评价指标(F1 - Measure,F_1)

$$F_1 = \begin{cases} C_p - S_p - 0.15\lg N_1 & \text{(构建单个文本的图模型时)} \\ C_p - S_p - 0.01\lg N & \text{(其他)} \end{cases} \tag{15}$$

式中 N_1——簇元素为 1 的数量

N——簇的数量

2.3.2　关键词抽取效果评价指标

第 1 类评价指标有 3 个：准确率（Precision，P）、召回率（Recall，R）和综合评价指标（F－Measure，F），公式为

$$P=\frac{X}{X+Y} \tag{16}$$

$$R=\frac{X}{X+Z} \tag{17}$$

$$F=\frac{2PR}{P+R} \tag{18}$$

式中　X——正确抽取到的关键词数

Y——错误抽取到的关键词数

Z——属于关键词但未被抽取到的词数

第 2 类评价指标为针对有序的关键词抽取结果的评价指标，包括平均倒数等级（Mean reciprocal rank，MRR）和二元偏好度量（Binary preference measure，Bpref）。其中，MRR 用来度量每个文档第 1 个被准确提取的关键词的排名情况，而 Bpref 则用来度量提取结果中错误提取的词语的排名情况，具体的计算公式为

$$M_{RR}=\frac{1}{|D|}\sum_{d\in D}\frac{1}{r_d} \tag{19}$$

$$B_{pref}=\frac{1}{|Q_1|}\sum_{r\in Q_1}\left(1-\frac{|F_0|}{|E|}\right) \tag{20}$$

式中　D——所有文档的集合

r_d——第 1 个正确提取结果的排序

Q_1——正确的关键词的集合

$|F_0|$——排列在正确提取词 $r\in Q_1$ 之前提取的错误词的数目

$|E|$——所有提取词的数目

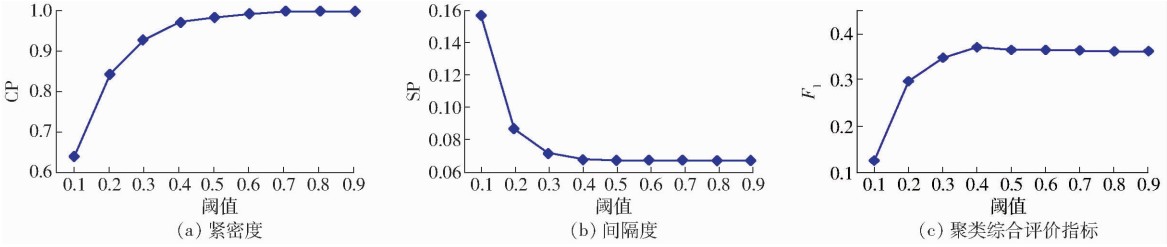
2.4　实验结果及分析

依据两两向量间的余弦相似度，对单个文本构建图模型时，需要设置合理的阈值。实验结果如表 1 所示。此时部分指标随阈值变化趋势如图 3 所示。

表 1　单个文本构建图模型阈值参数

Tab. 1　Parameters of single text built graph model

阈值	CP	SP	N_1	F_1
0.1	0.641 2	0.153 6	257	0.126 1
0.2	0.845 0	0.085 5	1 209	0.297 1
0.3	0.929 7	0.071 2	2 564	0.347 2
0.4	0.973 3	0.067 8	3 679	0.370 6
0.5	0.982 6	0.067 0	4 712	0.364 6
0.6	0.993 3	0.067 0	5 598	0.364 1
0.7	0.999 0	0.067 1	6 313	0.361 9
0.8	1	0.067 1	6 487	0.361 1
0.9	1	0.067 1	6 499	0.361 0



(a) 紧密度　(b) 间隔度　(c) 聚类综合评价指标

图 3　各指标变化趋势（单个文本）

Fig. 3　Trend of each index (single text)

由表 1 及图 3 可以看出，当阈值大于等于 0.4 时，CP 和 SP 逐步趋于稳定，且当阈值等于 0.4，F_1 最大，说明聚类效果最好。因此，单个文本构建图模型阈值参数设置为 0.4。

单个文本形成稳定的图结构后，要对所有的图进行归并聚类形成初始簇，需要设置合理的阈值。实验结果如表 2 所示。此时部分指标随阈值变化趋势如图 4 所示。

由表 2 及图 4 可以看出，当阈值在 0.5 ~ 0.7 之间时，CP 趋于稳定；虽然 SP 呈递增趋势，但要综合 CP 来设定阈值参数。阈值为 0.5 时，CP 为 0.929 0，SP 为 0.050 2，此时簇的数量 N 为 1 194，F_1 值最大，为 0.848 0。因此，归并聚类时阈值参数设置为 0.5。

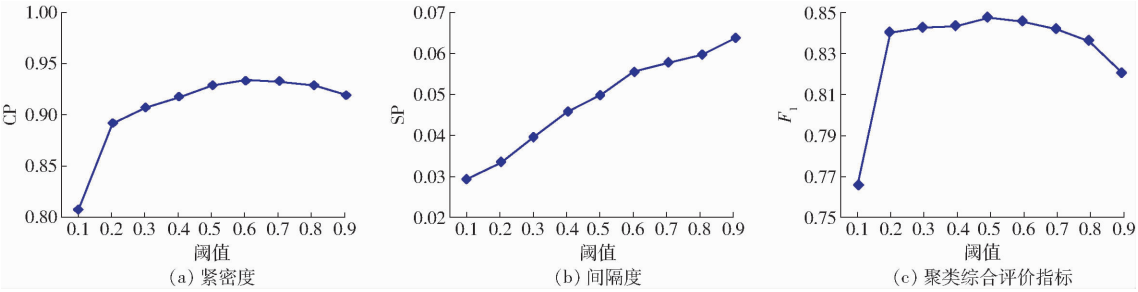
对簇进行过滤时，要进行品质评价。此时需要讨论参数 λ_1 和 λ_2，将 λ_1 在 0.1 ~ 0.9 的取值分别记为序号 1 ~ 9，实验结果如表 3 所示。

表 2　图结构归并聚类阈值参数

Tab. 2　Parameters of merged cluster of graph structure

阈值	CP	SP	N	F_1
0.1	0.806 1	0.029 5	11	0.766 2
0.2	0.891 8	0.033 3	66	0.840 3
0.3	0.907 6	0.039 8	280	0.843 3
0.4	0.917 6	0.045 8	662	0.843 6
0.5	0.929 0	0.050 2	1 194	0.848 0
0.6	0.933 8	0.055 8	1 618	0.845 9
0.7	0.933 1	0.057 9	1 979	0.842 2
0.8	0.929 4	0.059 8	2 278	0.836 0
0.9	0.919 4	0.064 0	2 716	0.821 1

由表 3 可以看出：当 λ_1 为 0.7，λ_2 为 0.3 时，CP 为 0.968 0，SP 为 0.057 2，此时簇的数量 N 为 234，说明能对簇进行有效过滤，此时 F_1 为 0.887 1，综



(a) 紧密度 (b) 间隔度 (c) 聚类综合评价指标

图 4 各指标变化趋势

Fig. 4 Trend of each index

表 3 品质评价公式参数

Tab. 3 Parameters of quality evaluation

序号	λ_1	λ_2	CP	SP	N	F_1
1	0.1	0.9	0.876 3	0.065 8	429	0.784 2
2	0.2	0.8	0.876 3	0.065 8	429	0.784 2
3	0.3	0.7	0.877 3	0.066 1	421	0.784 9
4	0.4	0.6	0.886 8	0.065 6	383	0.795 4
5	0.5	0.5	0.914 1	0.063 6	320	0.825 4
6	0.6	0.4	0.946 3	0.058 9	270	0.863 1
7	0.7	0.3	0.968 0	0.057 2	234	0.887 1
8	0.8	0.2	0.952 6	0.066 4	94	0.866 5
9	0.9	0.1	0.963 2	0.069 1	83	0.875 0

合评价最好。因此,归并聚类时阈值参数 λ_1、λ_2 分别设置为 0.7、0.3。

为了验证本文过滤方法的有效性,将过滤前的状态记为状态 1,采用本文过滤方法过滤后的状态记为状态 2,并与文献[34]提出的基于聚类显著程度的定量过滤指标对簇进行过滤的方法作对比,并记为状态 3,对比实验采用上述评价指标,结果如表 4 所示,结果说明簇品质公式能有效对簇的品质进行评价,本文过滤方法是行之有效的。同时对最终簇的元素来源文章数进行了统计,来源文章数规模最大为 21,最小为 4。

表 4 簇数量统计

Tab. 4 Numbers of clusters

状态	CP	SP	N	F_1
1	0.929 0	0.050 2	1 194	0.848 0
2	0.968 0	0.057 2	234	0.887 1
3	0.945 0	0.055 6	646	0.861 3

对簇进行信息类型标注,部分标注结果如表 5 所示。

表 5 部分标注结果

Tab. 5 Part of results

信息类型	词语
进出口	义乌 义乌市 内贸 口岸 货物
会议	第二十 四次会议 条例 人大常委会 全国政协 港澳台侨 会议 电视电话会议 第十四次 省人大 修改稿 修改意见 常委会 草案 政协 提案
展览	第十七届 中国花卉协会 博览会 海峡两岸 花卉 交易会 第三届 中国林学会 圆满结束 万里行 国际博览 博览会 本届 展会 开幕式 第八届 展区 观展 会展中心
地区	永清县 邯郸县 保定市 石家庄 唐山市 廊坊市 京冀 三河市
机构	工会主席 直属机关 劳动 工会 班组长 劳动法 员工 上岗 工资待遇 城镇职工 工资收入 职工 全总 住房补贴 安置
金融	授信 银行 信贷资金 贴息 小额 贷款 贴息贷款 抵押 分行 农行 银行业 人民银行 信用贷款 信贷 保险 准备金 投保 保险公司 理赔 定损 赔付 参保 监管部门 保监局 查勘 保险机构 被保险人
影视	戏剧界 话剧 舞台 剧本 电影 讲述 山丹丹花 首映 首映式 影片 纪录片 科技苑 电视 中央电视台 林海雪原 雪原
环境	二氧化硫 负氧离子 空气 立方厘米 减排
商家	餐馆 悬挂 店铺 饭店 招牌 商家
年代	七十年代 解放前 记载 数百年
出版	主编 出版 编纂 出版物
大学	华东师范大学 复旦大学 农业大学 农大

为了进一步验证本文方法在抽取关键词方面的有效性,将 TF - IDF、TextRank 以及文献[10 - 11, 35 - 36]中相应的关键词抽取方法分别作为对比实验。将上述模型按提及次序分别记为模型 1 ~ 6,将本文方法记为模型 7。实验数据为已进行关键词标注的 400 篇林业文本。实验结果如表 6 所示。

通过实验结果可以看出,本文所提方法在 MRR、Bpref、准确率和综合评价指标上均取得了最好的效果,在召回率方面取得了较好的效果,说明本文所提关键词抽取公式具有很好的关键词抽取

表 6　对比实验结果

Tab. 6　Results of comparative experiments

模型	*P*	*R*	*F*	MRR	Bpref
1	0.684 2	0.535 5	0.600 8	0.614 0	0.710 6
2	0.638 8	0.498 0	0.559 7	0.579 9	0.708 4
3	0.688 3	0.557 0	0.615 7	0.619 5	0.708 6
4	0.733 3	0.626 7	0.675 8	0.625 0	0.766 7
5	0.729 0	0.616 9	0.668 3	0.625 4	0.720 7
6	0.666 7	0.535 5	0.594 0	0.596 7	0.693 1
7	0.782 0	0.601 7	0.680 1	0.650 5	0.781 7

能力。

2.5　测试

2.5.1　测试流程

为了进一步验证本文所提出的关键信息抽取方法的有效性，开展了相关的测试实验工作，测试流程如图 5 所示。

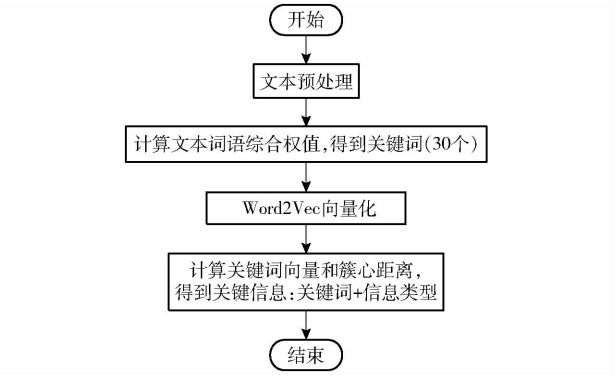


图 5　测试流程图

Fig. 5　Test process

具体步骤如下：①文本预处理，对文本进行分词、去停用词等操作。②依据关键词抽取公式，抽取权重排名前 30 的词语。③对抽取的关键词进行向量化表征。④通过计算比较关键词向量和各簇心之间的距离，得到关键词、最相似簇、最大相似度的三元组，根据最大相似度对 30 个三元组进行降序排序，取前 10 个。以最相似簇的标注类型作为词的信息类型，最终得到文本的关键信息：关键词 + 信息类型。

2.5.2　测试实例

选取一篇新的与林业政策新闻相关的文章，文章的部分内容如图 6 所示。

2.5.3　测试结果

抽取最大相似度排名前 10 的词语，结果如表 7 所示。

坚持把林业安全生产工作抓实抓细_重要信息坚持把林业安全生产工作抓实抓细

中国绿色时报 1 月 19 日报道（记者张兴国通讯员孙友）

1 月 16 日，国家林业局召开全国林业安全生产电视电话会议，深入学习习近平总书记等中央领导同志关于安全生产工作的系列重要指示批示，传达贯彻落实国务院安全生产委员会全体会议和全国安全生产电视电话会议精神，部署今后一个时期林业安全生产工作。1 月 12 日，国家林业局党组会议专门研究了林业安全生产工作。国家林业局党组书记、局长张建龙强调，安全问题事关重大，不能有任何侥幸麻痹思想，要继续抓紧落实。各地各司局单位要切实增强红线意识，按照管行业必须管安全、管业务必须管安全、管生产经营必须管安全的要求切实抓好林业安全生产工作，确保不出现重特大生产安全事故。国家林业局党组成员、副局长李树铭出席会议，对做好林业安全生产工作提出 5 点要求：一要不断提高认识。各地各有关司局单位要克服林业部门多年未出现大的生产安全事故的麻痹大意思想，进一步深化认识，认真履行好林业部门职责。二要强化责任落实。按照中央提出的党政同责、一岗双责、失职追责和管行业必须管安全、管业务必须管安全、管生产经营必须管安全的工作要求，各地要落实好林业安全生产属地管理和行业管理责任，国家林业局各有关司局单位要落实好本司局单位业务范围内的安全生产工作监管责任。要督促各类林业生产经营单位进一步落实好安全生产主体责任。三要加强源头治理。加大对重点领域、重点部位、重点环节的管控，不断加大对国有林区、国有林场、森林公园、湿地公园、沙漠公园等林区和林业生产经营单位的隐患排查治理工作，严防风险演变、隐患升级为生产安全事故。四要健全工作机制。各地和国家林业局有关司局单位要把抓好业务范围内的安全生产工作纳入重要议事日程。各地要进一步明确林业安全生产工作机构，加强队伍建设。加强林业生产安全事故应急救援预案建设。严格落实林业生产经营单位安全教育培训制度。五要搞好协调配合。各相关司局单位要按照管业务必须管安全的要求，抓好各自业务范围内安全生产管理工作。各地要认真

图 6　测试文章的部分内容

Fig. 6　Part of test article

表 7　测试结果

Tab. 7　Test results

类别	结果
机构	国家林业局、国务院、安全生产委员会
政策	林业安全生产、责任、安全
会议	电话会议、党组会议
林业	安全事故、林业生产

结果表明本文所提方法能从“关键词 + 信息类型”两部分表示文本关键信息，内容表述基本清晰，且具有很好的代表性和可读性。

3　结束语

为了使抽取的关键词综合特征明显，通过综合考虑词长、词跨度、标题等特征，将计算的综合权值作为词语特征值，通过构建融合词语特征、引入边权重的图模型对 TextRank 算法进行改进。经迭代收敛、归并聚类得到稳定的簇集合，对其过滤得到高品质的词语信息类型集合。实验表明，本文方法相对于其他关键词抽取方法具有更高的关键词抽取能力，最终形成的信息类型集合在紧密性、间隔性、综合评价指标上均表现良好。随机针对一篇林业政策新闻类的文本进行测试，结果表明，从“关键词 + 信息类型”两部分考虑，本文方法能有效提取出该文本的关键信息，说明本文提出的基于改进 TextRank 和簇过滤的林业文本关键信息抽取方法是有效的。

参　考　文　献

[1]　李思阳. 基于林业主题的 PageRank 算法优化的研究[D]. 哈尔滨：东北林业大学，2017.
LI Siyang. Research of PageRank algorithm optimization based on forestry theme[D]. Habin: Northeast Forestry University, 2017. (in Chinese)

[2]　ABDELHAQ H, GERTZ M, ARMITI A. Efficient online extraction of keywords for localized events in Twitter[J]. GeoInformatica, 2017, 21(2):365 – 388.

[3]　赵京胜，张丽，肖娜. 基于复杂网络的中文文本关键词提取研究[J]. 青岛理工大学学报，2018,39(3):106 – 112.

ZHAO Jingsheng, ZHANG Li, XIAO Na. Research on the Chinese text keyword extraction based on complex network[J]. Journal of Qingdao University of Technology, 2018, 39(3): 106 – 112. (in Chinese)
[4] 段青玲,张璐,刘怡然,等. 基于农业网络信息分类的热词自动提取方法[J/OL]. 农业机械学报,2018,49(7):167 – 174.
DUAN Qingling, ZHANG Lu, LIU Yiran, et al. Automatic extraction method of hot words based on agricultural network information classification[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(7): 167 – 174. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 20180720&journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2018.07.020. (in Chinese)
[5] CHEN Y, ZHOU R, ZHU W, et al. Mining patent knowledge for automatic keyword extraction[J]. Journal of Computer Research & Development, 2016, 53(8): 1740 – 1752.
[6] 陶洁. 基于新闻文本的关键词提取[D]. 武汉:华中师范大学,2019.
TAO Jie. Keywords extraction based on news text[D]. Wuhan: Central China Normal University, 2019. (in Chinese)
[7] 常耀成,张宇翔,王红,等. 特征驱动的关键词提取算法综述[J]. 软件学报,2018, 29(7):224 – 248.
CHANG Yaocheng, ZHANG Yuxiang, WANG Hong, et al. Features oriented survey of state-of-the-art keyphrase extraction algorithms[J]. Journal of Software, 2018, 29(7): 224 – 248. (in Chinese)
[8] WAN L. Extraction algorithm of english text summarization for English teaching[C]// Xiamen: 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2018.
[9] 文秀贤,徐健. 基于用户评论的商品特征提取及特征价格研究[J]. 数据分析与知识发现, 2019, 3(7):42 – 51.
WEN Xiuxian, XU Jian. Research on product characteristics extraction and hedonic price based on user comments [J]. Data Analysis and Knowledge Discovery, 2019, 3(7): 42 – 51. (in Chinese)
[10] 徐立. 基于加权 TextRank 的文本关键词提取方法[J]. 计算机科学, 2019, 46(增刊 1):142 – 145.
XU Li. Text keyword extraction method based on weighted TextRank[J]. Computer Science, 2019, 46(Supp. 1): 142 – 145. (in Chinese)
[11] 王涛,李明. 改进的关键词提取算法研究[J]. 重庆师范大学学报(自然科学版), 2019, 1(3):98 – 104.
WANG Tao, LI Ming. Study on an improved keyword extraction algorithm[J]. Journal of Chongqing Normal University (Natural Science), 2019, 1(3): 98 – 104. (in Chinese)
[12] 陶兴,张向先,郭顺利. 基于 DPCA 的社会化问答社区用户生成答案知识聚合与主题发现服务研究[J]. 情报理论与实践, 2019, 42(6):94 – 98, 87.
TAO Xing, ZHANG Xiangxian, GUO Shunli. Research of user-generated-answer knowledge aggregation and topic discovery service in social Q&A community based on DPCA[J]. Information Studies: Theory & Application, 2019, 42(6): 94 – 98, 87. (in Chinese)
[13] 何旭峰, 陈岭, 陈根才,等. 基于 LDA 主题模型的分布式信息检索集合选择方法[J]. 中文信息学报, 2017, 31(3):125 – 133.
HE Xufeng, CHEN Ling, CHEN Gencai, et al. A LDA topic model based collection selection method for distributed information retrieval[J]. Journal of Chinese Information Processing, 2017, 31(3): 125 – 133. (in Chinese)
[14] 胡琪,郝晓燕,张兴忠,等. 关键词抽取策略研究[J]. 太原理工大学学报, 2016, 47(2):228 – 232.
HU Qi, HAO Xiaoyan, ZHANG Xingzhong, et al. Research on the strategy of keyword extraction[J]. Journal of Taiyuan University of Technology, 2016, 47(2): 228 – 232. (in Chinese)
[15] 郭建波,谢飞. 基于多特征的关键词抽取算法[J]. 合肥工业大学学报(自然科学版), 2015, 38(9):1215 – 1219.
GUO Jianbo, XIE Fei. Keyphrase extraction algorithm based on multiple features[J]. Journal of Hefei University of Technology (Natural Science), 2015, 38(9): 1215 – 1219. (in Chinese)
[16] 朱煜松. 互联网舆情主题抽取方法研究[D]. 成都:电子科技大学,2018.
ZHU Yusong. Research on topic extraction method for internet public opinion[D]. Chengdu: University of Electronic Science and Technology of China, 2018. (in Chinese)
[17] 孙明珠,马静,钱玲飞. 基于文档主题结构和词图迭代的关键词抽取方法研究[J]. 数据分析与知识发现, 2019, 3(8):68 – 76.
SUN Mingzhu, MA Jing, QIAN Lingfei. Extracting keywords based on topic structure and word diagram iteration[J]. Data Analysis and Knowledge Discovery, 2019, 3(8): 68 – 76. (in Chinese)
[18] 塔瑞克. 基于图模型的关键词抽取研究[D]. 北京:北京邮电大学,2019.
TARIQUE Khan. Keyword extraction using a graph-based approach [D]. Beijing: Beijing University of Posts and Telecommunications, 2019. (in Chinese)
[19] 王斌,郭剑毅,线岩团,等. 融合多特征的基于远程监督的中文领域实体关系抽取[J]. 模式识别与人工智能, 2019, 32(2):133 – 143.
WANG Bin, GUO Jianyu, XIAN Yantuan, et al. Entity relations extraction in Chinese domain based on distant supervision with multi-feature fusion[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(2): 133 – 143. (in Chinese)
[20] 罗燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法[J]. 计算机应用, 2016, 36(3):718 – 725.
LUO Yan, ZHAO Shuliang, LI Xiaochao, et al. Text keyword extraction method based on word frequency statwastics[J]. Journal of Computer Applications, 2016, 36(3): 718 – 725. (in Chinese)
[21] STERCKX L, CARAGEA C, DEMEESTER T, et al. Supervwased keyphrase extraction as positive unlabeled learning[C]// ACL, 2016.
[22] 赵京胜,朱巧明,周国栋,等. 自动关键词抽取研究综述[J]. 软件学报, 2017, 28(9):2431 – 2449.
ZHAO Jingsheng, ZHU Qiaoming, ZHOU Guodong, et al. Review of research in automatic keyword extraction [J]. Journal of Software, 2017, 28(9): 2431 – 2449. (in Chinese)
[23] 李华灿. 基于统计与协同过滤的关键词提取研究[D]. 西安:西安电子科技大学,2015.
LI Huacan. Keyword extraction base on statwastical and collaborative filtering[D]. Xian: Xidian University, 2015. (in Chinese)
[24] 李畅,王永良,冯晓洁,等. 作战文书关键信息抽取方法[J]. 兵工自动化, 2011, 30(5):26 – 29.
LI Chang, WANG Yongliang, FENG Xiaojie, et al. Key information extraction from operational document[J]. Ordnance Industry Automation, 2011, 30(5): 26 – 29. (in Chinese)
[25] 刘志强,都云程,施水才. 基于改进的隐马尔科夫模型的网页新闻关键信息抽取[J]. 数据分析与知识发现, 2019, 3(3):120 – 128.
LIU Zhiqiang, DU Yuncheng, SHI Shuicai. Extraction of key information in web news based on improved hidden Markov model [J]. Data Analysis and Knowledge Discovery, 2019, 3(3): 120 – 128. (in Chinese)
[26] YANG Y, AGARWAL O, TAR C, et al. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction[C]// Proceedings of the 2019 Conference of the North, Minneapolis, 2019.

(下转第 172 页)

版), 2010, 36(2): 175 - 180.
LIN Fenfang, DENG Jingsong, DING Xiaodong, et al. Preliminary research on field measurement of spectral emissivity of rice in thermal infrared[J]. Journal of Zhejiang University(Agriculture and Life Sciences), 2010,36(2):175 - 180. (in Chinese)
[24] 王康丽, 韩迎春, 雷亚平, 等. 利用机载红外相机监测脱叶剂对棉花冠层温度的影响[J]. 中国棉花, 2018,45(10): 16 - 21.
WANG Kangli, HAN Yingchun, LEI Yaping, et al. Using airborne thermal infrared camera to monitor the effect of defoliant on cotton canopy temperature[J]. China Cotton, 2018,45(10):16 - 21. (in Chinese)
[25] 李毅, 张嘉娇, 杜波, 等. 水稻与褐飞虱化学关系的研究进展[J]. 植物生理学报, 2018,54(4):528 - 538.
LI Yi, ZHANG Jiajiao, DU Bo, et al. Research progress of chemical interactions between rice and brown planthopper[J]. Plant Physiology Journal, 2018,54(4):528 - 538. (in Chinese)
[26] WANG Liming, QIU Guoyu, ZHANG Xiying, et al. Application of a new method to evaluate crop water stress index[J]. Irrigation Science, 2005,24(1):49 - 54.
[27] MING Han, ZHANG Huihui, DEJONGE K C, et al. Estimating maize water stress by standard deviation of canopy temperature in thermal imagery[J]. Agricultural Water Management, 2016,177:400 - 409.
[28] 张智韬, 边江, 韩文霆, 等. 无人机热红外图像计算冠层温度特征数诊断棉花水分胁迫[J]. 农业工程学报, 2018, 34(15):77 - 84.
ZHANG Zhitao, BIAN Jiang, HAN Wenting, et al. Cotton moisture stress diagnosis based on canopy temperature characteristics calculated from UAV thermal infrared image[J]. Transactions of the CSAE, 2018, 34(15): 77 - 84. (in Chinese)
[29] GONZÁLEZ-DUGO M P, MORAN M S, MATEOS L, et al. Canopy temperature variability as an indicator of crop water stress severity[J]. Irrigation Science, 2006,24(4):233.
[30] 袁德栉. 不同灌水下限对水稻生长特性及产量的影响[J]. 广东农业科学, 2014,41(20):9 - 14.
YUAN Dezhi. Effects of irrigation threshold on growth characteristics and yield of rice[J]. Guangdong Agricultural Sciences, 2014,41(20):9 - 14. (in Chinese)
[31] 徐赛, 陆华忠, 周志艳, 等. 基于理化指标和电子鼻的果园荔枝成熟度识别方法[J/OL]. 农业机械学报, 2015, 46(12):226 - 232.
XU Sai, LU Huazhong, ZHOU Zhiyan, et al. Identification of litchi's maturing stage in orchard based on physicochemical indexes and electronic nose[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015,46(12):226 - 232. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20151230&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2015.12.030. (in Chinese)

(上接第 214 页)

[27] 王雪梅,陈兴蜀,王海舟,等.基于标签和分块特征的新闻网页关键信息自动抽取[J].山东大学学报(理学版),2019, 54(3):67 - 74.
WANG Xuemei,CHEN Xingshu,WANG Haizhou,et al. Automatic extraction of key information for news web pages based on tag and block features[J]. Journal of Shandong University(Natural Science), 2019,54(3):67 - 74. (in Chinese)
[28] 顾亦然,许梦馨.基于 PageRank 的新闻关键词提取算法[J].电子科技大学学报, 2017,46(5):777 - 783.
GU Yiran,XU Mengxin. Keyword extraction from news articles based on PageRank algorithm[J]. Journal of University of Electronic Science and Technology of China,2017,46(5):777 - 783. (in Chinese)
[29] 李航,唐超兰,杨贤,等.融合多特征的 TextRank 关键词抽取方法[J].情报杂志, 2017,36(8):187 - 191.
LI Hang,TANG Chaolan,YANG Xian,et al. TextRank keyword extraction based on multi feature fusion[J]. Journal of Intelligence,2017,36(8):187 - 191. (in Chinese)
[30] 李钰曼,陈志泊,许福.基于 KACC 模型的文本分类研究[J].数据分析与知识发现, 2019,3(10):89 - 97.
LI Yuman,CHEN Zhibo,XU Fu. Classifying texts with KACC model [J]. Data Analysis and Knowledge Discovery,2019, 3(10):89 - 97. (in Chinese)
[31] MANALU S R, WILLY, SUNDJAJA A M, et al. Review assessment support in open journal system using TextRank[C]// Journal of Physics Conference Series,Kobe,2017.
[32] 邓松,万常选.基于主题与概率模型的非合作深网数据源选择[J].软件学报, 2017, 28(12): 3241 - 3256.
DENG Song, WAN Changxuan. Non-cooperative deep web data source selection based on subject and probability model[J]. Journal of Software, 2017, 28(12): 3241 - 3256. (in Chinese)
[33] 彭圳生,巩青歌,高志强,等.基于密度及文本特征的新闻标题抽取算法[J].中文信息学报, 2018, 32(10):78 - 86.
PENG Zhensheng,GONG Qingge,GAO Zhiqiang,et al. Newstitle extraction algorithm based on density and text-features[J]. Journal of Chinese Information Processing, 2018, 32(10): 78 - 86. (in Chinese)
[34] 周冰,李飞,侯位昭,等.基于簇过滤的优势集模糊聚类集成[J].计算机与网络, 2019, 45(7):67 - 70.
ZHOU Bing,LI Fei,HOU Weizhao,et al. Dominant set fuzzy clustering ensemble based on cluster filtering[J]. Computer & Network,2019,45(7):67 - 70. (in Chinese)
[35] FLORESCU C, CARAGEA C. A position-biased PageRank algorithm for keyphrase extraction[C]// Proceedings of the 31st American Association for Artificial Intelligence,Hawaii,2017.
[36] LIU Z, HUANG W, ZHENG Y,et al. Automatic keyphrase extraction via topic decomposition[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,Cambridge,2010.